AUTOMATIC ADAPTATION OF A VOCAL TRACT MODEL

Blaise Potard and Yves Laprie

Speech Team, Loria / Lorraine University
LORIA, Campus Scientifique - BP 239, 54506 Vandoeuvre-ls-Nancy Cedex, France
email: {potard,laprie}@loria.fr

ABSTRACT

In this paper we present a method for adapting an articulatory model to a new speaker from acoustic data only. The main goal of this method is to make acoustic-to-articulatory inversion a fully automatic process. Speaker-specificity is modeled by a two dimensional scale factor, which makes it more flexible than VTLN methods.

Validation of the method is performed on three speakers by comparing the scale factors found on medical images of the vocal tract and those estimated. These experiments show that the method is accurate provided that the speech material covers a wide acoustic space.

Two additional experiments on longitudinal acoustic data are presented, in order to study vocal tract evolution with age.

Introduction

When performing acoustic-to-articulatory inversion, one of the usual preliminary requirements is to find a model of the speaker's vocal tract. When some articulatory data is available for the speaker, such as a sagittal X-ray view, or a mid-saggittal MRI image of the vocal tract, this task is usually easy to perform.

For "pure" acoutic-to-articulatory inversion however, only acoustic information is available from the speaker, which makes the task more complicated.

Several speaker adaptation methods have been presented in previous studies; the largest class of speaker adapation techniques are "vocal tract-length normalisation", commonly used e.g. when computing MFCC[2]. The vocal tract "adaptation" is fully automatic, but however quite crude, since it is assumed that most of the speaker variability can be represented by one scale factor.

More physiologically accurate models usually use two scale factors: one for the length of the oral tract, and a second one for the length of pharyngeal tract. Some articulatory models such as that of Maeda[6] thus include two scale factors, to roughly adjust the model to the speaker.

Some methods for adapting Maeda's articulatory model to the speaker from the acoustics only were proposed in some earlier works [9, 4]. These methods were however not fully automatic since they required some segmentation of speech, and phone-specific articulatory modelling. Although the segmentation into phones could arguably be done automatically using speech recognition techniques, the method still required some language specific adaptation, because it relied on the hypothesis that some phones were pronounced with quasi-identical articulatory configurations among all speakers of the language.

In this paper, we present a fully automatic method that does not require any manual intervention or modelling. It relies solely on the hypothesis that speakers aim at minimising the energy spent during speech production[3]. It is an extension of the method presented by the same authors in [10].

1. METHODOLOGY

Our approach to articulatory-inversion can be classified as analysis-by-synthesis, using an articulatory model coupled to a synthesiser. Our most recently published work[10] is based on variational calculus, and aims at finding an articulatory trajectory that minimises both acoustic error and articulatory effort.

The method we present here relies on the same framework, but unlike our previous method which uses only the 7 articulatory parameters of Maeda's articulatory model (cf. Fig. 1), our new method uses 9 parameters, i.e. the 7 articulatory parameters, plus two additional parameters corresponding to the scale factors of the oral and pharyngeal cavities.

1.1 Vocal tract scale factors

The two vocal tract scale factors were introduced by Maeda in his original model[7]. To simplify, the effect of applying an oral scale factor of l_1 is to multiply all dimensions in the oral tract by this factor, the effect of a pharyngeal scale factor l_2 is to multiply all dimensions in the pharyngeal tract by l_2.

In order to have two additional components P_8 and P_9 that behave similarly to the other components $P_1 \ldots P_7$ of Maeda's articulatory model, we linearly transformed Maeda's scale factors so as to get "normalised" parameters. In effect, the transformation applied is the following:

$$P_8 = (l_2 - 1) * 10, P_9 = (l_1 - 1) * 10.$$

The $[-3:3]$ interval (which is the usual interval of variation allowed for the other articulatory model parameters) thus corresponds to the scale factors interval $[0.7:1.3]$, which should cover the majority of adult vocal tract shapes dimensions.

1.2 Global variational calculus and cost function

In [10], the iteration conducted was the following:

$$-\left[\frac{\partial f_j}{\partial \alpha_i^\tau}(t)\right]^{-1}_{i=1..M, j=1..N} (F(t) - f(\alpha^\tau(t))) = -\lambda \vec{m}\alpha''^\tau(t) + \beta \vec{k}\alpha^\tau(t) + \gamma \frac{\partial \alpha^\tau}{\partial \tau}(t) \quad (1)$$

where:

- t is a time index over the speech sequence,
- $\alpha(t)$ is an articulatory vector (of dimension M),

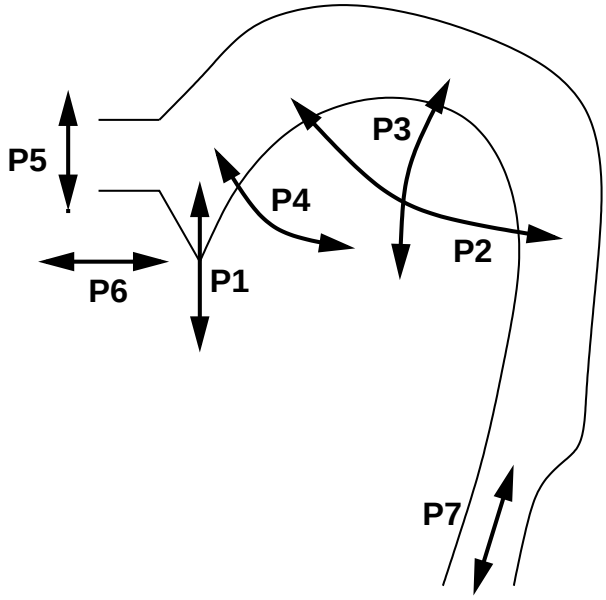

Figure 1: Maeda's articulatory model and control parameters.

- f is the articulatory-to-acoustic mapping that we wish to invert,
- $F(t)$ is an acoustic vector (of dimension N),
- τ is the index of the iteration,
- λ is a weight factor for the articulatory smoothness,
- β is a weight factor for the articulatory position,
- γ is a factor to control the articulatory distance between two iterations,
- $\vec{m}$ and $\vec{k}$ are weighting vectors of dimension M, to apply different weight factors to individual components of the articulatory vector.

In [10], M was equal to 7, and the components of $\vec{m}$ and $\vec{k}$ were all equal to 1. In the current study, M is modified to 9, and the scale coefficients are not taken into account in the articulatory cost function, i.e. the two last components of $\vec{m}$ and $\vec{k}$ are equal to 0.

Furthermore, we introduced the additional constraint that α_8 has an unique value along the whole trajectory (i.e. for all t), and as well for α_9.

1.3 Inversion procedure

The idea of this method is to introduce the scale coefficients in an acoustic-to-articulatory procedure. The iteration presented above converges in most cases towards the articulatory trajectory that can produce the measured acoustic signal with the minimum articulatory cost. It is assumed here that a given acoustic signal will always be "harder" (i.e. with a higher articulatory cost) to be produced identically by a different speaker.

Since variational calculus is guaranteed to converge towards an optimal extremum – provided the initial solution is close enough to that extremum – and since our method has been shown to usually converge towards the original articulatory trajectory with a 7-dimensional articulatory vector – even with an arbitrary initial articulatory trajectory – we expect this method to be fairly reliable to find the optimal scaling factors, provided there are few errors in the acoustic vectors used as input.

sent.	l_1	l_2
PB01	0.99	0.99
PB02	0.98	1.02
PB03	0.98	1.03
PB08	0.98	1.01
PB09	0.98	1.00
PB15	1.01	0.98
PB17	1.00	1.00
PB18	0.97	0.99
PB24	0.95	0.99
PB28	1.01	0.98
Avg.	0.98 ± 0.02	1.00 ± 0.02

Table 1: Scale factors for speaker PB

2. EXPERIMENTS AND RESULTS

The validity of the method was checked on a few speakers for which we could compute the correct scale factors from articulatory data.

We ran the inversion procedure on the original "PB" speaker that provided the articulatory data use to build Maeda's model, and on two male speakers from the ASPI European project[8].

Additionnaly, experiments were conducted on two speakers (one male, one female), for which we had obtained audio recordings at various ages. This tests the validity of the method in the sense that recordings from comparable years should lead to similar scale factors, but this also allows us to investigate a recurring question in vocal aging studies, on whether a pattern can be observed in the evolution of the vocal tract dimensions along age.

In all these experiments, the acoustic features used were the three first formants frequencies, tracked automatically using Wavesurfer or WinSnoori. In one case, the formant extraction was done manually. The use of more sophisticated acoustic features such as LPC coefficients or MFCC is not appropriate in this particular case, since our articulatory model only allows us to generate the vocal tract transfer function. Formants frequencies are still to our knowledge the only reliably extractable features that will have a close match in the transfer function. MFCC for example will incorporate the spectral tilt due to the source / lips radiation impedance and will therefore not match the transfer function.

2.1 Reference speakers and validation

The first speaker we ran experiment on is the reference speaker used to build the articulatory model[6], PB[1]. The scale factors for this reference speaker are both 1. The speech signal being very noisy, the formant tracking was done manually. Speech inversion with speaker adaptation was conducted on all 10 sentences of the corpus.

The second and third speakers were speakers YL and FH from the ASPI European project[8]. Reference scale factors were measured on an MRI image in the case of YL, on an X-ray image for speaker FH.

Table 1 shows the scale factors found through inversion for PB. Independent experiments were conducted on each sentence of the corpus, which are about 2 seconds long. For

Speaker	meas. scale fact.	est. scale. fact.
YL	1.15, 1.20	1.15, 1.19
FH	1.19, 1.07	1.16, 1.05

Table 2: Scale factors for speakers YL and FH

this speaker, we see that the scale factors are always very close to the expected values $(1.00, 1.00)$. The method thus seems to be quite successful on that particular speaker, even with rather short speech sequences. However, the result may be slightly biased in that case, since the articulatory model deformation modes are adapted to the speaker, which may lead to a better scale factors discrimination than for other speakers. Additionnally, the sentences of the corpus are phonetically balanced, which may also make the task easier since they cover a wide acoustic space, and therefore leave less leverage for speaker variability.

To further validate the method, we thus conducted similar experiments on the two male French speakers of the ASPI project. To measure the phrayngeal and oral scale factors, we superimposed Maeda's grid on an image of the midsagittal slice of the vocal tract. The scale factors were then adjusted to get the best visual fit between the model and the vocal tract image.

Table 2 gives the scales factors for speakers YL and FH, found on medical images (column "meas. scale fact.") and estimated by inversion (column "est. scale. fact."). For both speakers, we see that the factors found through our inversion procedure are very similar to those measured. In this experiment, we inverted a sequence of logatomes in case of YL, and a phonetically balanced sentence for FH. The procedure was also applied on a single VCV or VV, or repetitions of a single VCV, but the results obtained were then largely incorrect, for both speakers. It seems that the acoustic space covered in the sequence has to be fairly large to yield significant results.

2.2 Vocal aging experiments

A fully automatic procedure was run on audio recordings of two native speakers of English, QE (female) and AC (male). The originality of these corpus are the very large time intervalle covered by the recordings: QE is partially represented from age 26 to 76, AC from age 38 to 95. These two corpus were generously provided by J. Harrington[5] for the VAE team of 2008 CLSP summer Workshop. Formants tracking was done with WinSnoori in the case of QE, with Wavesurfer in the case of AC.

For each recording, acoustic-to-articulatory inversion with speaker adaptation was performed. The corresponding pharyngeal and oral scales are plotted on Fig. 2 and Fig. 3. Note that some years have several recordings, and some others none at all.

The observations of these figures show a clear trend for the pharyngeal scale: it is increasing with age in the case of QE, more slowly increasing with age until about 82 then decreasing in the case of AC.

Regarding the oral scale, it seems to be slowly decreasing with age for QE, very slowly increasing until age 85, then decresing fast for AC.

The results found for the pharyngeal scale show the expected trend. Furthermore, plotting the evolution of the average fundamental frequency along age show that the pharyngeal and fundamental frequency are strongly correlated (cf. Fig 4).

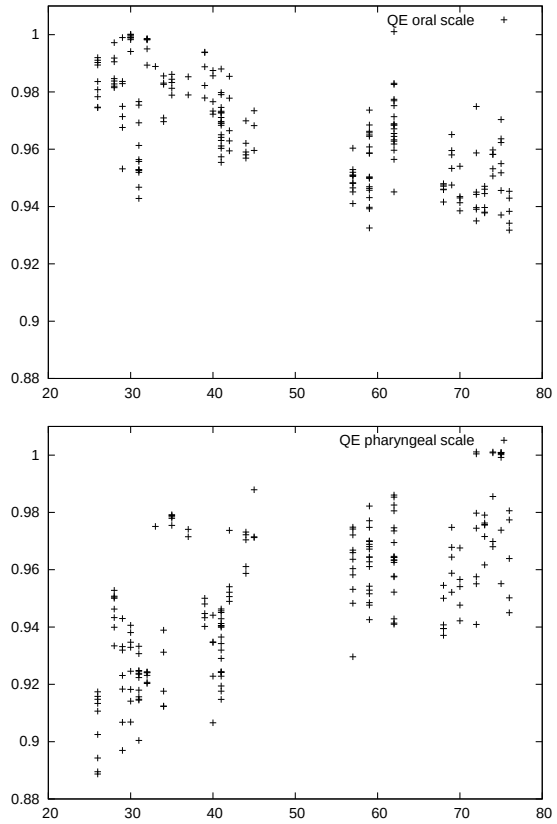

Figure 2: Oral and pharyngeal scales factors evolution with age for speaker QE.

The evolution of the oral scale is harder to interpret. Since the palate is a rigid body, we were not expecting much variation in this area (although some variations could possibly be explained by a change in the amount of lips protrusion with aging). Regardless, the amount of variation observed is small compared to the pharyngeal scale, below the margin of error observed for speaker PB, and the two speakers show opposite trends. We can thus consider that no notable evolution can be observed.

3. CONCLUSION

These preliminary experiments prove that our method is quite reliable for determining the correct scale factors for several speakers, although the margin of error is still larger than when fitting the model from articulatory data. We observed that the acoustic space has to be quite large for accurate results. A small phonetically-balanced sentence is usually enough to adapt the articulatory model to the speaker. Using this method, it is thus possible to build an inversion system that adapts itself to the speaker fully automatically and with very little speech material.

This method also appears to be accurate enough to allow us to observe the pharyngeal lengthening associated with aging.

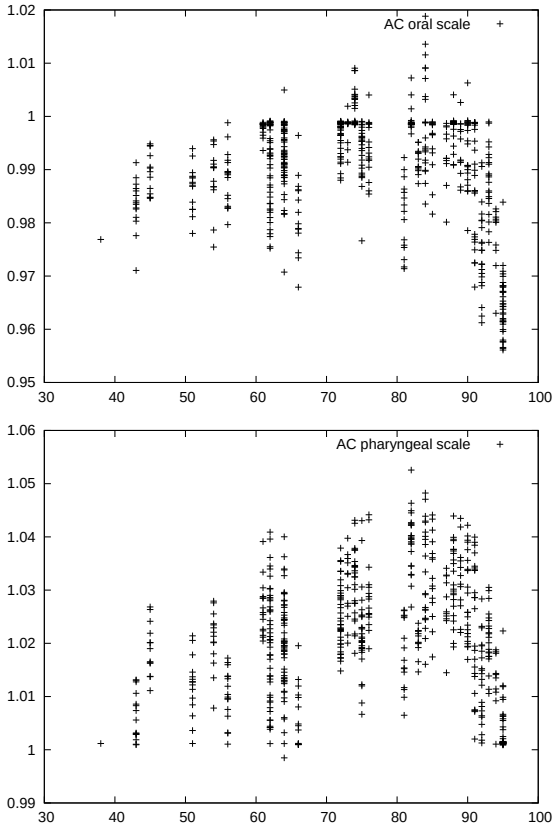

Figure 3: Evolution of oral and pharyngeal scales factors with age, for speaker AC.

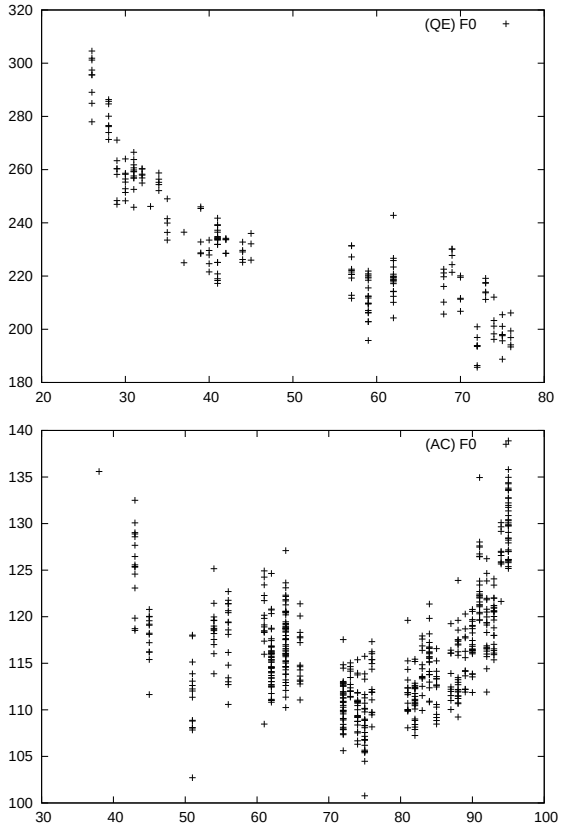

Figure 4: Evolution of fundamental frequency (in Hertz) with age, for speaker QE and AC.

Acknowledgement

The authors would like to express their gratitude to J. Harrington for providing the longitudinal acoustic data, and to all the members of the VAE team of the 2008 CLSP summer workshop: P. Beyerlein, A. Cassidy, V. Kholhatkar, E. Lasarcyk, S. Shum, Y. C. Song, W. Spiegl, G. Stemmer, P. Xu, and more specially E. Nöth

REFERENCES

[1] A. Bothorel, P. Simon, F. Wioland, and J.-P. Zerling. *Cinéradiographies des voyelles et consonnes du Français*. Travaux de l'institut de Phonétique de Strasbourg, 1986.

[2] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. ICASSP*, Atlanta, GA, May 1996.

[3] G. Fant. *Acoustic Theory of Speech Production*. The Hague: Mouton & Co., 1960.

[4] A. Galván-Rodrigez. *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives*. Thèse de l'Institut National Polytechnique de Grenoble, 1997.

[5] J. Harrington1, S. Palethorpe, and C. I. Watson. Age-related changes in fundamental frequency formants: a longitudinal study of four speakers. In *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007.

[6] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

[7] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

[8] S. Maeda, M.-O. Berger, O. Engwall, Y. Laprie, P. Maragos, B. Potard, and J. Schoentgen. Acoustic-to-articulatoy inversion: Methods and acquisition of articulatory data. Technical report, ASPI Consortium, November 2006.

[9] M. Naito, L. Deng, and Y. Sagisaka. Model-based speaker normalization methods for speech recognition. In *Proc. EUROSPEECH*, Budapest, September 1999.

[10] B. Potard and Y. Laprie. A robust variational method for the acoustic-to-articulatory problem. In *Interspeech, Brighton*, Sept. 2009.